RESEARCH STATEMENT

By Guanqun Cao

Michigan State University

My research interests focus mainly on statistical methodologies with real life applications, including but not limited to medical research, acoustics, food science and Geography. In particular, my work centers on three areas: 1) **non- and semi-parametric modeling of complex data** (functional/longitudinal data and high-dimensional data), 2) **hypotheses testing for non-identifiable models**, and 3) **nonparametric modeling for diffusion tensor imaging data**.

I am interested in developing novel and powerful statistical methodology, establishing deep statistical theory, and analyzing complex real data. The right balance of theoretical, computational and applied skills enables me to continue application-driven and theoretically challenging research. The research undertaken in my thesis, together with future research plans, are described below.

1 Statistical inference on non- and semi-parametric models for complex data

Simultaneous inference for mean & derivative functions in functional data analysis

Functional data analysis (FDA) has recently become a very hot topic in statistical research, as recent technological progress in measuring devices now allows one to observe spatiotemporal phenomena on arbitrarily fine grids, that is, almost in a continuous manner. This area remains distinct due to its contribution to climatology, medicine, meteorology, economics, etc. Characterizing nonlinear variation in *FDA* is a challenging problem. It provides the opportunity for statisticians to develop new methodologies to address it. In particular, when random curves are observed on regular dense grids, the existing literature on *FDA* focused on pointwise estimation and inference. This, however, is not sufficient to provide understanding of the variability of the estimator of the whole regression curve and its derivative, nor can it be used to correctly answer questions about the curve's shape.

Cao, Yang and Todem (2011) established asymptotic correctness of the proposed simultaneous confidence band for mean functions using various properties of spline smoothing. The spline estimator of mean curves and the accompanying confidence bands are asymptotically the same as if all the random curves are recorded over the entire interval, without measurement errors. Computational efficiency is a huge advantage for the spline methods in analyzing large data sets and in performing simulation studies. This work is illustrated by questions from food science, namely, whether different contents of fat in meat have different absorbance on spectrum. We not only provide an estimator of the difference of the two mean functions but also apply the aforementioned confidence bands to such an estimator. Whereas the motivation and one application of this work is to the food industry, the resulting methodology has broad applications. Recently, I have been collaborating with geographic scientists (*Center for Global Change and Earth Observations at Michigan State University*), who are eager to apply this novel method to classification of the agriculture product regions using a remote sensing data sets.

An immediate goal after focusing on mean curves is to extend theoretical findings for the derivative of mean curves, since estimation and inference of derivatives of the mean functions in *FDA* are of equal importance. For example, in economics, consistent and direct estimation of derivatives are essential for the estimation of elasticities, returns to scale, substitution rates and average derivatives. In **Cao**, Wang, Wang, and Todem (2011), we first show that the proposed estimators of derivatives of the mean curves are semiparametrically

efficient by taking advantage of strong approximation and Karhunen-Loève representation. Moreover, we establish consistency results for derivatives of covariance functions and their eigenfunctions in *FDA*.

Simultaneous inference for covariance functions in functional data analysis

In *FDA*, covariance estimation plays a critical role in functional principal component analysis, functional generalized linear models, and other functional nonlinear models. There has been some recent work on nonparametric covariance estimation in functional data, which is mostly based on kernel smoothing. However, large data set problems, including for example, recorded speeches for voice recognition and electroencephalogram data, require computationally more efficient methods. **Cao**, Wang, Li and Yang (2011) considered nonparametric estimation of the covariance function using reduced rank tensor product B-splines. Specifically, based on the asymptotic distribution of the maximum deviation of the estimator, we proposed a new simultaneous confidence envelope for the covariance function, which can be used to visualize the variability of the covariance estimator and to make global inferences on the shape and other properties of the true covariance. Consider for example the hypothesis that the covariance is stationary. In a speech recognition application, the classic functional linear discriminant analysis assumes that the random curves from different classes share a common covariance function. To justify this assumption, we further extend our confidence envelope method to a two-sample problem, where one can test whether the covariance functions from two groups are different.

Inference of change-point in single index models

Single index models are more flexible and less restrictive than parametric models, as the link function is not specified. Moreover, compared with nonparametric models, single index models avoid the curse of dimensionality by virtue of linear combination of variables. In the application of regression methods, ignoring possible change points may result in serious errors in drawing inference about the process under study. Based on the density-weighted average derivative estimation method, **Cao**, Wang, Wu and Zhao (2008) proposed a statistic to test whether a change point exists or not in single index models by taking advantage of U-statistics and Brownian bridges. The key idea needed to obtain the null distribution of the test statistic is to apply a permutation technique. We rigorously showed that the permuted statistic has the same distribution in asymptotics under both null and alternative hypotheses.

2 Hypotheses testing for non-identifiable models

Many statistical models arising in applications contain non-identified parameters. A more practical illustration of this problem comes from the missing data literature, where non-identifiable models are frequently encountered. Due to identifiability concerns, tests concerning the parameters of interest may not be able to use conventional theory and it may not be clear how to assess statistical significance. Essentially, standard estimation and inference techniques may fail due to the models being overparameterized. **Cao**, Todem, Yang and Fine (2010) derived the limiting distribution of the test statistic and proposed theoretically justified resampling approaches to approximate its asymptotic distribution. The methodology's practical utility is illustrated in simulations and an analysis of quality-of-life outcomes from a longitudinal study on breast cancer (*International Breast Cancer Study Group*).

3 Nonparametric modeling for diffusion tensor imaging data

Another interest of mine is diffusion tensor imaging (DTI), which is a popular in vivo brain imaging technique that combines magnetic resonance imaging (MRI) technology with diffusion measurements of water molecules in order to produce neural tract images. DTI has a tremendous impact on brain function studies, as it is the only approach available to track brain white matter fibers and living microstructures noninvasively. In this research area, one models microstructures in soft tissues by integral curves, which are not observed directly. Instead, for example a two-dimensional vector field is observed on a regular grid perturbed by additive random noise. The object of interest is an estimator of an integral curve driven by the vector field starting at a fixed location. Currently, the only theoretically rigorous works focus on kernel regression estimation for the vector fields and plug-in estimation for corresponding integral curves.

In **Cao**, Sakhanenko, Yang and Carmichael (2011), we have collaborated with a neuroscientist from the *University of California at Davis*, and have constructed a B-spline estimator of the vector field and a plug-in estimator of the integral curve. The properties of tensor product splines help us build an estimator which is asymptotically normal. More precisely, the properly normalized difference between the estimated and true curves converges weakly to a centered Gaussian process with a certain covariance function. Then confidence ellipses along the integral curve are constructed. As an alternative to a kernel based approach, our estimator has no asymptotic bias and corresponding confidence ellipses can be computed faster.

4 Further directions

My sight is focused on several exciting challenges. For instance, the simultaneous confidence bands/envelopes constructed by polynomial spline open the door to FDA. To bring this methodology to a wider audience, we must expand it, so it would work with random sparse grids and irregularly observed grids. Another direction of my interest in FDA is registration or warping of the functional data. Typical variation for functional data is phase variation, that is, not all features of the curves occur at the same locations. For example, human growth is a complex sequence of hormonal events that do not happen at the same rate for each child. Hence, comparing the intensity of the pubertal growth spurt of two children should be taken at their respective ages of peak velocity instead of any fixed age (Ramsay and Silverman, 2005). To this end, registration of the functional data not only can be adopted as a necessary data preprocessing step for curve regression or simultaneous inference, but also could be recognized as a possible source of information for clustering and classification.

I have a general interest in FDA classification. For these large-scale data, classifying observations into different functional groups is a first step in order to gain more sophisticated knowledge of different functions. Many existing classification analysis methods belong to the general framework of multivariate analysis. A more efficient way to look at such data is to incorporate the information that is inherent in the time order and the smoothness of processes over time, i.e., FDA. The two-sample test for covariance functions in **Cao**, Wang, Li and Yang (2011) opens up the opportunity for functional discriminate analysis. I would like to explore supervised and unsupervised curves classification based on FDA by borrowing the ideas of support vector machines and neural networks..

In addition, numerous interesting issues with respect to diffusion tensor imaging data remain to be explored. In **Cao**, Sakhanenko, Yang and Carmichael (2011), the tensor product spline estimator has been studied only for the two-dimensional vector field. I plan to work on extending the estimator to more general cases, such as high dimensional vector fields, as diffusion is truly a three-dimensional process.

Ultimately, I see the tight coupling of statistical theory and practical, goal-directed applications. I am looking forward to taking on these exciting challenges.

References

[1] **Cao, G.**, Li, Y., Wang, L. and Yang, L. (2011) Spline confidence envelopes for covariance function in dense functional/longitudinal data. *Under review.*

[2] **Cao, G.**, Sakhanenko, L., Yang, L. and Carmichael, O. (2011) Spline estimation of integral curves from noisy vector field data. *Manuscript.*

[3] **Cao, G.**, Todem, D., Yang, L. and Fine, J. (2010) Evaluating statistical hypotheses using non-identifiable estimating functions. *Under review.*

[4] **Cao, G.**, Wang, L., Wang, J. and Todem, D. (2011) Spline Confidence Bands for Functional Derivatives. *Journal of Statistical Planning and Inference*, revision submitted

[5] **Cao, G.**, Wang, Z., Wu, Y. and Zhao, L. (2008) Inference of change-point in single index models. *Science in China Series A: Mathematics*, Vol. 51, No. 10, 1855-1870.

[6] **Cao, G.**, Yang, L. and Todem, D. (2011) Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics*, forthcoming.

[7] Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis.* Springer, New York.